Federal Republic of Germany

Patent Application (unexamined)
DE 39 29 481 A 1

Int. Cl. [5]:
G 10 L 3/00

File Number: P 39 29 481.1
Application Date: 5 September 89
Layed Open: 15 March 90

German Patent Office

Convention Priority:
7 September 88 JP 63-222309
Applicant:
Hitachi, Ltd., Tokyo, JP
Inventors:
Ichikawa, Akira, Musashino, Tokyo, JP;
Asakawa, Yoshiaki, Kawasaki, Kanagawa, JP;
Amano, Akio, Higashimurayama, Tokyo, JP;
Hataoka, Nobuo, Pittsburgh, Pa., US
Representatives:
Strehl, P., Dipl.-Ing., Dipl.-Wirtsch.-Ing.;
Schübel-Hopf, U., Dipl.-Chem., Dr.rer.nat.;
Groening, H., Dipl.-Ing., Patent Attorneys;
Schulz, R., Dipl.-Phys. Dr.rer.nat.,
Patent Attorney and Counselor at Law, 8000 München

Request for examination submitted pursuant to § 44 PatG

Process and device for preprocessing voice signals

The voice data filter used for the processing and the device for preprocessing voice signals comprises a number of microphones (103, 104, 105), which are disposed spatially apart from one another. The sound acting on the microphones is converted in the succeeding A/D converters (106, 107, 108) into a digital serial signals, which forms an input signal for a neural network. The neural network excludes background noises wherein partially data are used which are obtained from the parallax information gained through the offset configuration of the microphones. The data obtained from the neural network are subsequently transmitted to a digital signal processor in order to filter out the noise.

Specification

The invention relates generally to signal processing and, in particular, to a process and a device for preprocessing voice signals in order to improve the signal-to-noise ratio in the voice signals supplied to a voice processor.

Several processes for improving the signal-to-noise ratio in voice signals are known, wherein the frequency properties of the noise are examined in advance in order to be able to subtract the noise components subsequently from the voice signal. These known processes, however, rest on the false assumption that the background noise is uniform. Such systems operate typically with two microphone inputs in order to be able to subtract the corresponding signals and to exclude thereby the background noise. The use of so-called neural networks has also been discussed already (Proceedings of ASJ (Acoustic Society of Japan), Spring Meeting, 3-p-13, pp. 253-294, May 1988).

The system known from the last mentioned publication shows improved capabilities and has a signal-to-noise ratio which is superior to earlier techniques, however, it was found that the comprehensibility is therein reduced.

The term "neural network" here includes two types of neural networks. In the first type the neural network comprises equivalent elements processing in parallel which are interconnected according to a dynamically self-organizing programming in a non-monitored manner, i.e. self-training, independently of whether or not a "trainer" is present. In the second type of neural network, the network comprises equivalent elements, processing in parallel, which are in advance fixedly connected with one another through training. Such a network can later no longer "learn" anything.

The human voice is output from the mouth as a sequence of densifications and dilutions of the air molecules. The voice-forming organs, through which the voice information is output, are different in each person. Through the physical differences

between individual people the physical properties of the voice signals deviate considerably from one another if they are considered as physical signals. In addition, from the diverse sound sources from diverse directions noises or background noises is generated. The discrepancies in the physical properties of voice signals therefore have no commonality.

The object of the invention is creating a process and a device with which voice data can be obtained with improved comprehensibility and clarity.

For a solution of this task in the process or the corresponding device according to the invention the signals of a number of transducers which transduce sound into electric signals, such as for example microphones, are used as input signals of a neural network. The voice conversation is carried out by human beings without difficulties, even at a high noise level, partially through the use of both ears. The number of microphones yields input information, such as parallax information, which the neural network can use in order to carry out a filtering of sound.

The neural network carries out a training process such that only physical properties which are common to the input signals from a number of microphones and a pure voice signal, which is supplied for training from the output side of the network, are transmitted. All other signals are filtered out. Consequently, only signals are transmitted which have exclusively the physical properties of the voice while the noise components are suppressed. The signal-to-noise ratio of the system is thereby significantly improved.

With the process according to the invention and the corresponding device it is possible to increase the signal-to-noise ratio of voice information acquired at a very high noise level. Consequently, the following advantages are obtained:

The reliability of the voice recognition is increased thereby that a filter according to the invention is disposed in front of a voice recognition device.

The reliability of the recognition is also increased thereby that the filter according to the invention is disposed in front of a voice coding device, whereby a coded voice with a high signal-to-noise ratio which is readily recognized, is obtained such that voice communication can be carried out even at high noise level.

The reliability of the recognition is also increased thereby that the filter according to the invention is disposed in front of one of the various customary types of voice analysis apparatus whereby it becomes possible to detect distortions of the voice at a high noise level (under such conditions a person generally raises his voice, in order to speak above the noise whereby the voice deviates from its customary form).

With the system according to the invention it is therefore possible to increase the signal-to-noise ratio in voice data without simultaneously impairing comprehensibility.

In the following, an embodiment example of the system according to the invention will be explained in further detail in conjunction with the drawing. Therein depict:

Fig. 1 the structure of a filter for improving the signal-to-noise ratio with a representation of the training process connected therewith,

Fig. 2 a device in which the filter of Figure 1 is used, and

Fig. 3 examples for use of the system according to the invention .

In Figure 1 the disposition of a filter for improving the signal-to-noise ratio comprising a neural network is depicted. Figure 2 shows the application of this filter and a filtering and training system.

The neural network shown in Figure 1 comprises a number of "neurons" which are disposed in a first to fourth layer 114, 117, 119 and 121. As is generally known, the individual neurons can be formed by processing units which carry out a valuation or weighting of the signals at their input or they can be emulated by a conventional von Neumann machine. It is understood that for structuring the network also more or fewer neurons and/or layers or planes than in Figure 1 can be used.

As shown in Figure 2, voice signals 101 and noise signals 102 which are input into a number of microphones 201 are digitized by a multiplexing A/D converter 202 and subsequently supplied to a switch 203. In train mode the digitized signal is conducted from switch 203 to an internal bus 204, and under the control of a microprocessor (µ CPU), 205 stored in a store 206 in order to built up the neural network according to the procedures contained in the microprocessor 205. The result of the training is obtained in the form of weighting factors for the connections between the elements of the neural network.

Each weighting factor determined thus is conducted via a signal line 207 to a digital signal processor (DSP) 208, which comprises a neural network for noise filtering. The digital signal processor 208 thus represents a neural network for noise filtering in which the weightings are already determined ("learned"). If the system is used as a noise filter, the input voice signal 101 (and the noise signal 102) is input directly via the microphones 201, the A/D converter 202 and the switch 203 into the signal processor 208 in order to obtain at the output of processor 208 a signal 209 with improved signal-to-noise ratio. If the configuration is only used as a noise filter, the elements required for training also do not need to be present.

The operational function of this noise filter and the training procedure will be described in conjunction with Figure 1. During the leaning, some of the parts shown in Figure 1 can be realized through virtual parts of the microprocessor 205 and of the store 206 of Figure 2, while in the execution of an actual operation only those parts are in the signal processor 208 which form the filter shown in Figure 1. It is understood that it is also possible that the microphones 201 and the A/D converter 202 are disposed at a different location and are connected via a digital line with the signal processor 208, which in this case forms alone the device.

For simplification the description takes place with reference to a configuration comprising two input systems. However, the configuration can also comprise in the same manner three or more input systems.

In the representation of Figure 1 it is assumed that the output signal of the qth neuron element in the pth layer is equal to $O_{p,q}$ and the output signal of the rth

element in the (p-1)th layer is equal to $O_{p-1,r}$. For simplification of the description it is further assumed that the transmission property between the input x and the output y is identical for all elements and is represented by

$$y = f(x) \quad (1)$$

Then the following applies:

$$I_{p,q} = \Sigma w_{p-1,q,r}(O)_{p-1,r} \quad (2)$$

$$O_{p,q} = f(I_{p,q}) \quad (3)$$

It can be seen from equation (1) that the processing contains many calculations for forming the sum of products which the signal processor carries out. The neural network includes preferably a large number of neuron elements which have the properties expressed by equation (3) and which are connected with one another in a hierarchical structure. It should be noted that, although the neural network shown in Figure 1 comprises four layers or planes, the number of layers is not necessarily limited to four.

The mixture of the voice signals 101 and the noise 102, which is supplied via the microphones 103 and 104 to the A/D converters 106 and 107, is converted there into digital signals which are routed further to shift registers 112 or 113. The shift registers 112 and 113 are provided, together with a shift register 124 (later explained in further detail) for the purpose of shifting sequentially the data synchronously to the sampling period of the A/D converters and to output data in each stage. The output signals of the various stages of the shift registers 112 and 113 are subsequently supplied to the elements 115 or 116 in the first (input) layer 114 of the neural network.

The output signals of elements 115 and 116 of the first layer are routed on the basis of the relationships expressed by the equations (2) and (3) to elements 118 of the second layer 117. The same applies to the connection between the elements 118 in the second layer 117 and the elements 120 in the third layer 119 as well as the connection between elements 120 in the third layer 119 and the elements 120 in the third layer 119 and the elements 122 in the fourth (output) layer 121. Through the processing of the signals in the elements on the basis of the relationships which are

represented by equations (2) and (3), signals 128 are output with an improved signal-to-noise ratio at the output terminals 123 of the output layer 121. If the output signal is taken out from one of the output terminals 123 as an external output signal, an output voice signal 209 (Fig. 2) with improved signal-to-noise ratio is obtained.

A description of the training process in the neural network forming the noise filter follows.

The backpropagation process which is known in the architecture of neural networks, is suitably applied for the training process in the present system. Such backpropagation process is described, for example in the literature M.I.T. Press, "Parallel Distributed Processing", Vol. 1(1986), Chapter 8, pp. 318-362.

The training process will now be explained with reference to Figure 1. For simplification some symbols will be introduced. The value of the output signal 128 of each element 122 in the output layer 121 is denoted by $O_{4,i}$, the value of the output signal of the jth element of the third layer 119 by $O_{3,j}$, the value of the output signal of the kth element in the second layer 117 by $O_{2,k}$ and a nominal output value, which is connected as a training input to the ith element in the fourth layer 121, by $T_{4,i}$. With respect to the error signal which is obtained for each signal in the course of the backpropagation, the value of the error signal for the ith element in the fourth layer 121 is denoted by $\delta_{4,i}$, the value of the error signal for the jth element in the third layer 119 by $\delta_{3,j}$ and the value of the error signal for the kth element in the second layer 117 by $\delta_{2,k}$. Furthermore, it is assumed that the transmission properties of the elements in all layers are identical and correspond to that expressed by equation (3). Moreover, let f' be the derivative of the function f. The connection factor between the ith element in the output layer 121 and the jth element in the third layer 119 is denoted by $w_{3,i,j}$, and the connection factor between the jth element in the third layer 119 and the kth element in the second layer 117 by $w_{2,j,k}$.

Various voice types 101 and various noise types 102 are entered separately into the microphones 103, 104 and 105 for the purpose of training. The signal input into microphone 105 comprises a pure voice signal, it is used for the nominal output

value $T_{4,i}$. The signals are stored in the particular stores 109, 110 and 111 (locations in store 206 of Figure 2). The stored voice and the stored noise are added in adders 129 and 130, in order to combine signals onto which a noise is superimposed. These signals are conducted to the shift registers 112 and 113. Data regarding the extent to which the noise is superimposed and the combination of voice and noise are repeatedly prepared for various expected states and are used as training input signals. In the implementation in practice the superposition is carried out by using an arithmetic function of the microprocessor 205 of Figure 2. The nominal output value $T_{4,i}$ is a voice signal which corresponds to the training input signal and defines the degree up to which the voice in the training input signal has been improved as a result of the improvement of the signal-to-noise ratio. The input microphone 105, the A/D converter 108 and the store 111 for the nominal output value $T_{4,i}$ can also be used for an input, which means as microphone 103 (or 104), D/A converter 106 (or 107) and store 109 (or 110), as is indicated by the connection line 127. The voice for the nominal output value $T_{4,i}$ is input into the shift register 124 and the output signals 125 from the various stages of the shift register 124 are input into the corresponding elements 122 in the output layer 121 of the neural network as nominal output signals 125.

If into each element in the first layer 114 is input a training input signal (voice and noise superimposed on one another), on the basis of the relationships expressed by equations (2) and (3), sequentially for each element from the input layer to the output layer an output signal is obtained. After the output signal for each element has been obtained, error signals are determined sequentially from the output layer 121 to the lower layers. The correction of the connection factors between the pth layer and the (p+1)th layer takes place using the error signals for the (p+1)the layer and the values of the output signals in the pth layer. In the following, for simplification only the process for the correction of the connection factors $w_{3,i,j}$ and $w_{2,j,k}$ will be explained. The corresponding process is repeated for the following layers down to the input layer.

For the correction of the connection factors $w_{3,i,j}$ and $w_{2,j,k}$ are required the

value $O_{2,k}$ of the output signal of the kth element in the second layer 117, the value $O_{3,j}$ of the output signal of the jth element in the third layer 119, the value $\delta_{3,i,j}$ of the error signal of the jth element in the third layer 119 and the value $\delta_{4,i}$ of the error signal of the ith element in the fourth (output) layer 121. The values for $O_{2,k}$ and $O_{3,j}$ can be obtained through a forward calculation by applying input signals to the first layer 114, as described above. The values $\delta_{4,i}$ and $\delta_{3,j}$ can be calculated from the following equations:

$$\delta_{4,i} = (T_{4,i} - O_{4,i}) f' \left(\sum_j w_{4,i,j}(O_{3,j})\right). \quad (4)$$

$$\delta_{3,j} = f' \left(\sum_k w_{2,j,k}(O_{2,k})\right)(\delta_{4,i})(w_{3,i,j}). \quad (5)$$

Next, $w_{3,i,j}$ and $w_{2,j,k}$ are corrected. If the correction values are expressed by $\Delta w_{3,i,j}$ and $\Delta w_{2,j,k}$, these correction values can be calculated as follows:

$$\Delta w_{3,i,j} = \alpha(\delta_{4,i})(O_{3,j}) \quad (6)$$

$$\Delta w_{2,k,j} = \alpha(\delta_{3,j})(O_{2,k}) \quad (7)$$

α can be set by experimentally checking the convergence speed. Equations (6) and (7) make possible the correction of all connection factors between the output layer and the third layer and between the third and the second layer. The connection factors between the second layer and the input layer can be corrected in the same way as the connection factors between the third and the second layer.

In this way, all connection factors are corrected once. With different input data and nominal values (such that differ from the above values with respect to the voice, the noise, the mutual levels and the mutual phase relation), the above process for correcting the connection factors is repeated. Every time the process is repeated, a weighting factor E is determined as follows:

$$E = 1/2 \sum_i (T_{4,i} - O_{4,i})2. \quad (8)$$

The weighting factors are averaged over all training patterns. If the mean

value is smaller than a preset threshold value, it is established that the training process is concluded.

If the location of a speaking person and the positions of the microphones are restricted to a preset region, the voice information for training is also input under the corresponding conditions and the internal voice signal is generated taking into account the levels and the phase differences between the microphones in this configuration. Thereby the effectiveness of the improvement of the signal-to-noise ratio is significantly increased. If for the location of the speaking person a certain region is to be permitted, the training input voice also corresponds to this region. The corresponding conditions can also be readily derived, for example through an internal synthesization on the basis of the foundation of acoustics (for example, it is sufficient to consider the delay of the voice signals which results from the spacing between the speaking person and the microphones, and the [inverse] square law attenuation).

It should be noted that it is also possible to subject the input signal to a complex Fourier transformation or the like and subsequently to input it, for example, in the frequency domain into the neural network. In such a case, the input layer can be provided for the frequency and the phase or for the real part and the imaginary part in two-dimensional form. The output can be an output signal in the frequency domain, which is transformed back into the wave form [time] domain. In this process one of the known domain projection transformations and a corresponding inverse transformation is required.

Some application of the above described filter are represented in Figure 3.

The voice recognition can be improved, for example, thereby that a noise filter 301 structured according to the above description is disposed in front of a voice recognition device 302, in order to obtain therefrom an improved output signal 303.

The noise filter 301 can also precede a voice coder device 304, whereby an encoded voice is obtained at its output 305, which can be readily recognized such that even at a very high noise level, a voice connection is possible.

Lastly, the noise filter 301 can also be disposed in front of a conventional voice

analysis apparatus 306, whereby it becomes possible to detect distortions of the voice at a high noise level if, for example, the voice is raised by a person in order to drown out the background noises, whereby the voice differs from its customary form.

The voice data filter used in the process or the device for the preprocessing of voice signals thus comprises a number of microphones which are disposed spaced apart from one another. The sound acting upon the microphone is converted in succeeding A/D converters into a digital serial signal which forms an input signal for a neural network. The neural network excludes the background noises, wherein partially data are used which are obtained from the parallax information which is obtained through the offset disposition of the microphones. The data obtained from the neural network, are subsequently transmitted to a digital signal processor in order to filter out the noise.

Patent Claims

1. Device for decreasing the noise in voice recognition systems, **characterized by**
   - a number of spatially separately disposed transducers (103, 104, 105; 201) for generating a number of electric voice signals which correspond to the sound acting upon the transducer,
   - a neural network with a number of layers (114, 117, 119, 121) wherein each layer comprises a number of neuron elements (115, 116; 118; 120; 122) and the layers include an input layer (114) and an output layer (121,
   - a first communication unit (106, 112) in order to transmit the electric voice signal from a first (103) of the transducers to each element (115) of a first set of neuron elements in the input layer, and by
   - a second communication unit (107, 113) in order to transmit the electric voice signal from a second (104) of the transducers to each element (116) of a second set of neuron elements in the input layer.

2. Device as claimed in claim 1, characterized in that each of the transducers (103, 104) comprises a unit for generating an analog electric voice signal which corresponds to the sound acting thereon, and the first and the second communication unit comprise each a unit (106; 107) for converting the analog electric voice signal into a first or second serial digital voice signal.

3. Device as claimed in claim 2, characterized in that the first and the second communication unit comprise each a shift register (112; 113) for changing the first or the second serial digital signals into a corresponding first or second series of output signals, wherein the output signals of the first and second series each form the input signal for a neuron element (115; 116) of the first and second set of the input layer (114).

4. Device as claimed in claim 3, characterized by a switch (203) for the selective application of a pure electric voice signal and a mixed electric voice/noise signal to the neural network and through a unit for carrying out a monitored training process in the neural network in agreement with the connected pure electric voice signal and the combined electric voice/noise signal whereby neuron weighting data are obtained which represent the transmission properties between the neuron elements of the neural network.

5. Device as claimed in claim 3, characterized by a unit (207) for transmitting the neural weighting data from the neural network to a digital signal processor, and by a digital signal processor (208) for processing combined voice/noise signals in agreement with the neural weighting data.

6. Device as claimed in claim 4, characterized by a unit for carrying out a Fourier transformation on at least one of the electric voice signals before it is transferred to the neural network.

7. Device as claimed in claim 4, characterized by

- a number N of additional spatially separated transducers for generating a number of electric voice signals, which correspond to a sound acting thereon, wherein N is a positive integer greater than zero,
- N additional communication units for transmitting the electric voice signal from each of the N additional transducers to each element of an Nth set of neuron elements in the input layer,
- wherein each of the additional N transducers comprises a unit for generating an analog electric voice signal corresponding to the sound acting thereon,
- wherein the additional N communication units comprise a unit for converting the analog electric voice signal into the particular Nth serial digitized voice signal, and

- wherein each of the additional N communication units comprises shift registers for changing the first or second serial digitized signals into the corresponding Nth series of output signals, wherein each output signal of the Nth series represents an input signal for a neuron element of the Nth set of the input layer.

8. Process for noise reduction in acoustic signals, characterized by the process steps:
(a) the receiving of sound waves from a number of positions,
(b) the generation of a number of electric sound signals corresponding to the sound waves from each of the positions,
(c) the transmission of the electric sound signals to one set of neuron elements (115; 116) in an input layer (114) of a neural network,
(d) the calculation of an output signal in the neural network which is derived from the electric sound signals at the first and second set of neurons, and
(e) the output of the output signals from an output layer (121) of the neurons of the neural network.

9. Process as claimed in claim 8, characterized in that the process step (b) comprises the generation of a number of analog electric voice signals and the conversion of the number of analog electric voice signals into a corresponding number of serial digitized voice signals; and that the process step (c) comprises the transmission of each of the digital voice signals to the corresponding set of neuron elements (115; 116) of the input layer (114) of the neural network.

10. Process as claimed in claim 10 [sic], characterized by the further process steps of selective application of a pure electric voice signal and a mixed electric voice/noise signal to the neural network and of carrying out a monitored training process in the neural network in agreement with the connected pure electric voice signal and the combined electric voice/noise signal, whereby the

neuron weighing data are obtained which represent the transmission properties between the neuron elements of the neural network.

11. Process as claimed in claim 10, characterized by the further process steps of transmission of the neural weighting data from the neural network to a digital signal processor (208), the transfer of mixed voice/noise signals from the neural network to the signal processor, and the processing of the combined voice/noise signals in the signal processor in agreement with the neural weighting data.

12. Process as claimed in claim 10, characterized by the further process step of carrying out a Fourier transformation on at least one of the electric voice signals before it is supplied to the neural network.

13. Device for reducing signal noise, characterized by

- a number of spatially separated transducers (103, 104, 105; 201) for generating a number of electric voice signals which correspond to a sound acting thereon,
- a number of digitizing units (106, 107, 108; 202) for converting analog mixed voice/noise signals into digital voice/noise signals,
- a unit (207) for transmitting digitally coded neural weighting data from a neural network to a digital signal processor,
- a digital signal processor (208) for processing the digital mixed voice/noise signals from the converters in agreement with the weighting data into filtered digital sound data, and by
- a unit for transmitting the filtered digital sound data to a corresponding receiving unit for the digital sound data.

14. Device as claimed in claim 13, characterized in that the receiving unit for the digital sound data comprises a unit for generating analog filtered sound data from the digital filtered sound data.

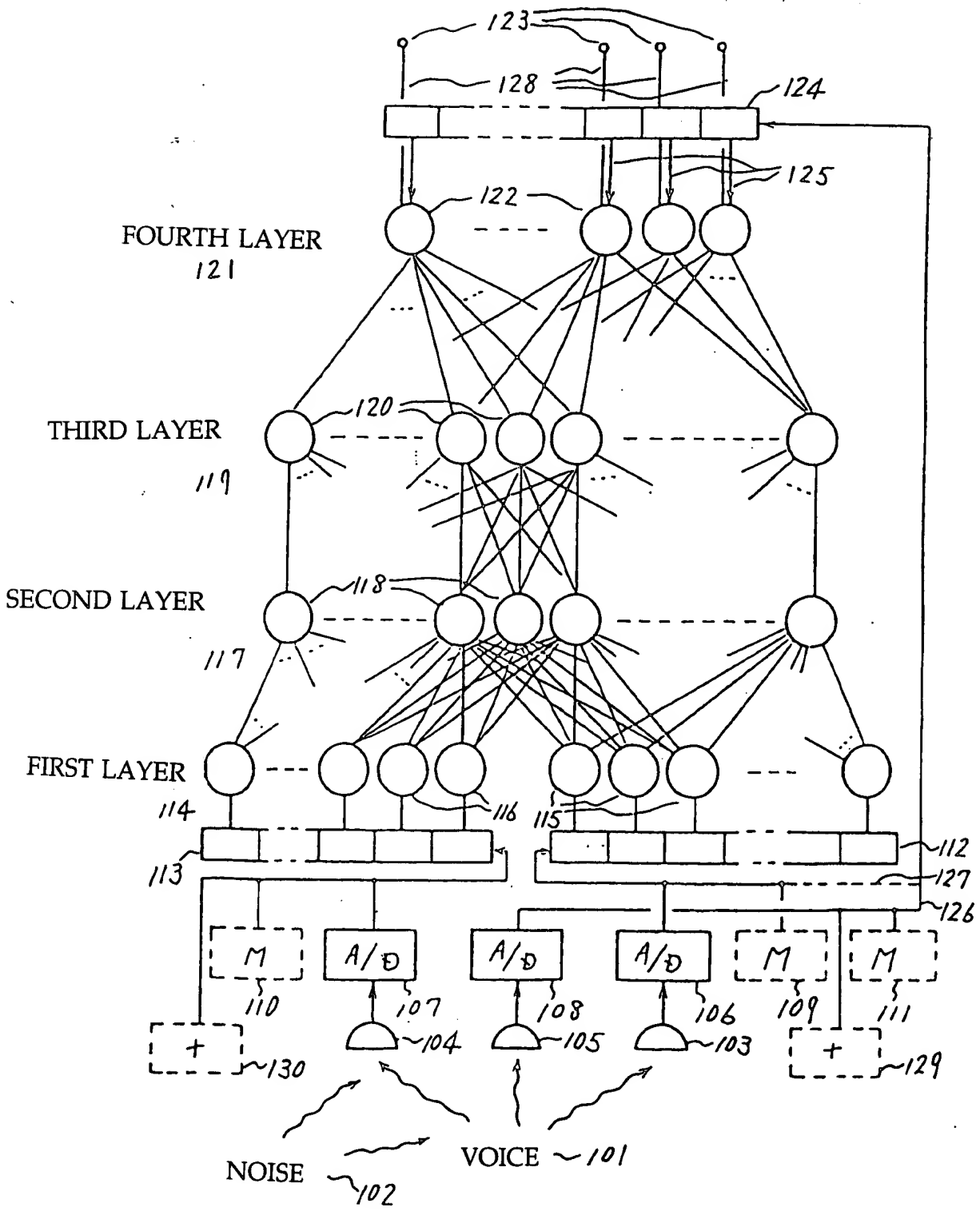
15. Device as claimed in claim 14, characterized by a loudspeaker for generating filtered sound waves from the analog filtered sound data.

---

2 sheets of drawings enclosed

Nummer: DE 39 29 481 A1
Int. Cl.[5]: G 10 L 3/00
Offenlegungstag: 15. März 1990

Fig. 1



Nummer: DE 39 29 481 A1
Int. Cl.[5]: G 10 L 3/00
Offenlegungstag: 15. März 1990

Fig. 2

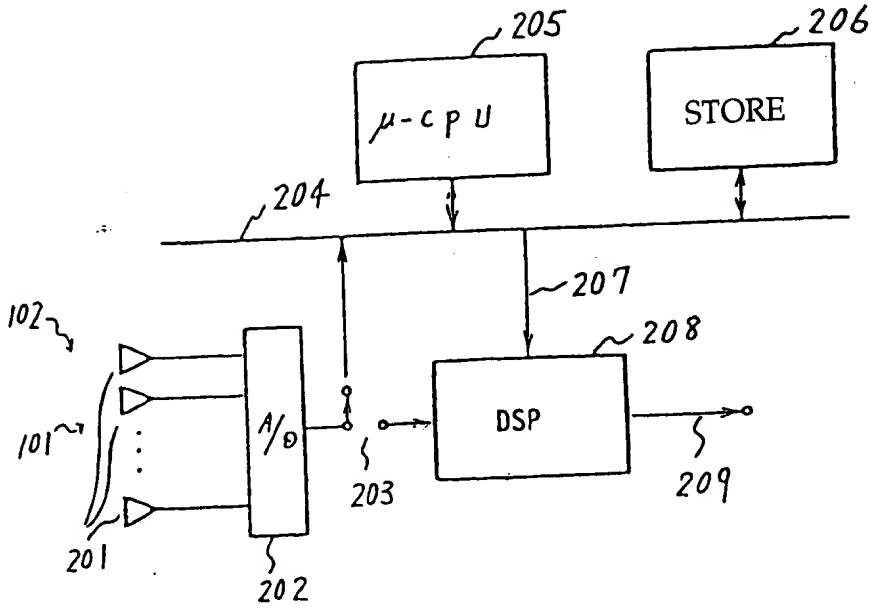


Fig. 3